Университет ИТМО

Проект

по дисциплине «Визуализация и моделирование»

Автор1: Литвак Игорь
Автор2: Сулейманов Руслан
Автор3: Штрейх Анна
Поток: ВИМ 1.1
Группа: К3241
Факультет: ИКТ
Преподаватель: Чернышева А.В.

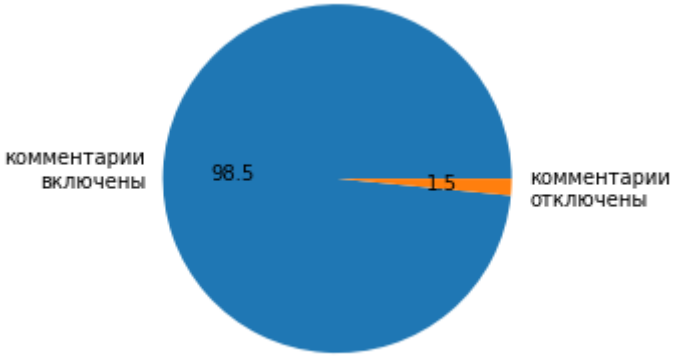
Санкт-Петербург, 2021 г.

Выбранный датасет — статистика видео из вкладки тренды видеохостинга YouTube (ссылка).

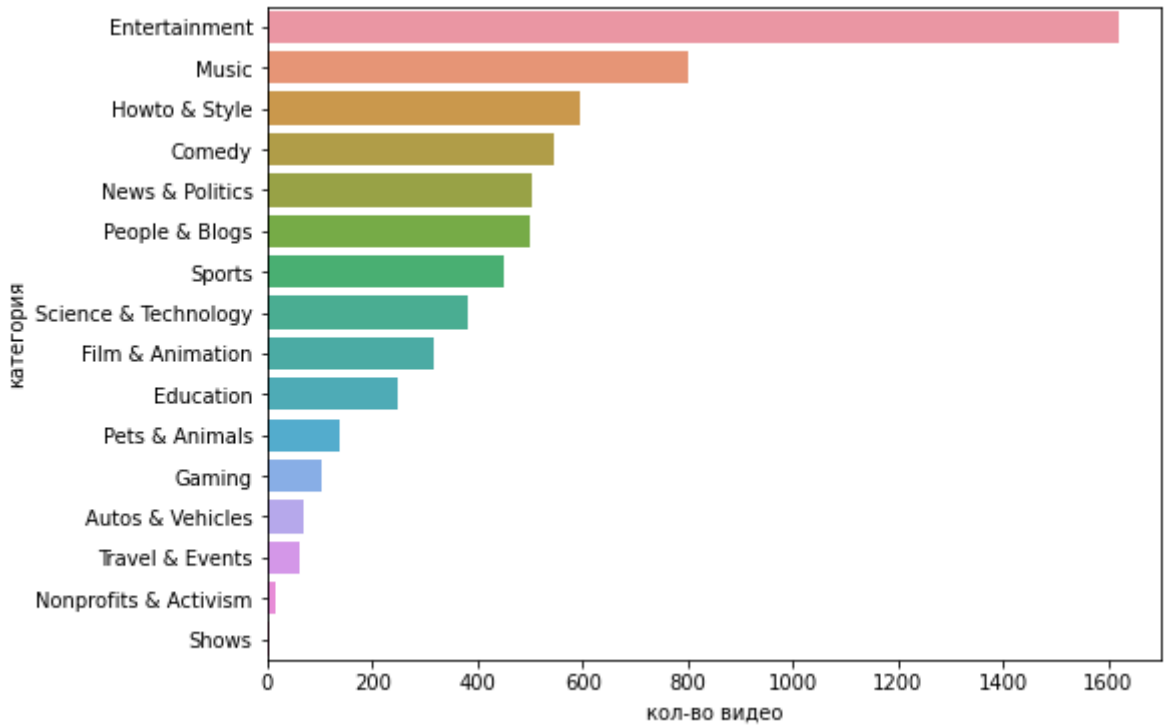
Data – storytailling

Как-то раз, один мудрый человек сказал: *“Скажи мне кто твой друг, и я скажу кто ты.”* Так и мы решили, узнать что за друзья у видеоролика ютуб, что б узнать о нем.

Давайте рассмотрим некоторое видео. Оно имеет свое название, дату тренда, число лайков, комментариев, дизлайков, теги и другие описывающие ее данные, конечно есть небольшой процент видеороликов, где этого нет: например 1,5процента видеороликов не имеют комментарии.

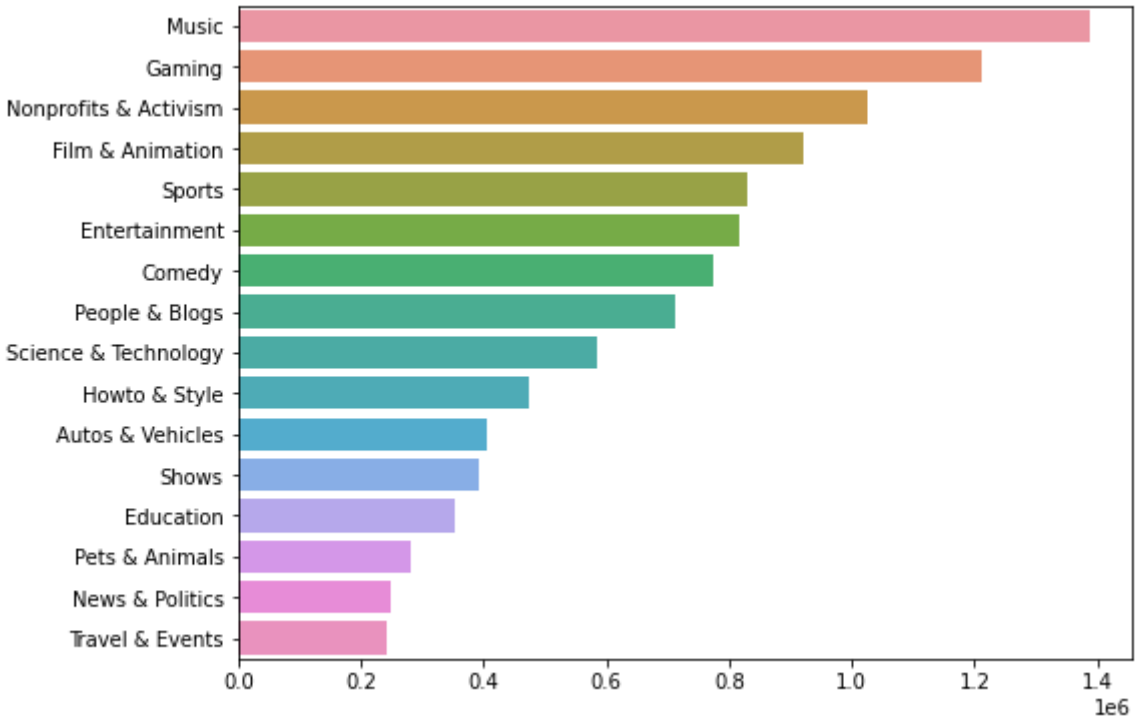


Но число их также мало, как и количество популярных видеороликов с категорией shows.



Так как же узнать, к какой категории относится видео? Да очень просто! Нужно всего-то сравнить все видеоролики, выискивая исключительные особенности для каждой категории, применяя различные алгоритмы, обучая машинку.

Давайте вернемся к данным, а для примера, именно, к музыке, как к самой популярной категории.



Из чего состоит музыкальный клип? Большое количество лайков, большое количество просмотров, быстрый выход в тренды, ну и название, в котором всегда есть определенный шаблон. Собственно, на последнее и стоит опираться. Сравнив все видеоролики из одной категории, мы вероятно сможем угадать категорию данного нам на вход видеоролика.

Так может и займемся этим?

Machine – learning

Задачей для машинного обучения с учителем выбрана классификация видео по категориям по их названию. В датасете каждому видео присвоена одна из 16 категорий, по которым и будет производиться классификация.

Для классификации были использованы два алгоритма: наивный Байесовский классификатор (далеко не лучший алгоритм, использовался для сравнения) и LinearSVC из библиотеки scikit-learn.

Перед прогоном данных через алгоритмы они были нормализованы - удалены все символы не являющиеся буквами английского алфавита, удалены единичные символы, произведена лемматизация, удалены слова из списка стоп-слов (предлоги и т.п.), удалены слова, которые встречаются меньше 5 раз на весь датасет и больше чем в 70% заголовков.

Для тренировки и тестирования 40000 записей в датасете были разделены в соотношении 80/20.

Результат:
Наивный классификатор Байеса показал общую точность в 69%, наименьшая в категории Nonprofits & Activism (0%), наибольшая в категории Music (84%)
Классификатор LinearSVC показал общую точность в 84%, наименьшая в категории People & Blogs (72%), наибольшая в категории Shows (100%)

Вывод: полученная точность является довольно высокой, учитывая относительно большое количество категорий (16), многие из которых сходны и короткий текст, с которым вынужден работать классификатор (у большинства видео это всего 5-10 слов).